Sanity Checks for Saliency Metrics

Richard Tomsett,[1][*] **Dan Harborne,**[2][*] **Supriyo Chakraborty,**[3] **Prudhvi Gurram,**[4] **Alun Preece**[2]

[1]Emerging Technology, IBM Research, Hursley, UK
[2]Crime and Security Research Institute, Cardiff University, Cardiff, UK
[3]IBM Research, Yorktown Heights, NY, USA
[4]Booz Allen Hamilton and CCDC Army Research Laboratory, Adelphi, MD, USA
rtomsett@uk.ibm.com, {harborned, preecead}@cardiff.ac.uk, supriyo@us.ibm.com, gurram_prudhvi@bah.com

Abstract

Saliency maps are a popular approach to creating post-hoc explanations of image classifier outputs. These methods produce estimates of the relevance of each pixel to the classification output score, which can be displayed as a saliency map that highlights important pixels. Despite a proliferation of such methods, little effort has been made to quantify how good these saliency maps are at capturing the true relevance of the pixels to the classifier output (i.e. their "fidelity"). We therefore investigate existing metrics for evaluating the fidelity of saliency methods (i.e. saliency metrics). We find that there is little consistency in the literature in how such metrics are calculated, and show that such inconsistencies can have a significant effect on the measured fidelity. Further, we apply measures of reliability developed in the psychometric testing literature to assess the consistency of saliency metrics when applied to individual saliency maps. Our results show that saliency metrics can be statistically unreliable and inconsistent, indicating that comparative rankings between saliency methods generated using such metrics can be untrustworthy.

Introduction

Despite their popularity, deep neural networks (DNNs) are widely acknowledged to lack interpretability: their complex internals, with thousands or millions of parameters and nonlinear elements, make it difficult to explain how and why a DNN maps inputs to outputs. This has led to many groups of researchers developing a variety of methods to help improve DNN interpretability. Saliency maps represent a popular class of explanation methods aimed at achieving this goal for DNNs operating on image data. They are designed to provide a measure of the relevance of each pixel to the DNN's output (Montavon, Samek, and Müller 2018). Many methods have been proposed in the recent past for generating saliency map type explanations (Simonyan, Vedaldi, and Zisserman 2013; Zeiler and Fergus 2014; Bach et al. 2015; Montavon et al. 2017; Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). However, most of this prior work focuses on developing new methods to generate the saliency maps, without much work evaluating the quality of these

methods. Different saliency methods provide different maps for the same image and corresponding DNN output, so human analysts need some way to decide which method provides the best estimates of the true pixel relevance values in order to choose between them.

Several authors have proposed axioms that explanation methods should adhere to in order to be produce appropriate explanations (Sundararajan, Taly, and Yan 2017; Kindermans et al. 2017). These can be assessed by inspecting the mathematics behind the explanation methods, and working out if the method is consistent with the axioms. However, it is quite possible for useless explanation methods to be consistent with currently proposed axioms. For example, assigning uniform relevance scores to every pixel produces an explanation that is conservative, continuous, and implementation invariant (Montavon 2019). To ensure that an explanation method actually assigns suitable relevance scores, we must assess whether the method truly discriminates appropriately between more and less relevant features in the input. This property has been referred to as the selectivity (Bach et al. 2015) or fidelity (Alvarez-Melis and Jaakkola 2018) of an explanation: how well it agrees with the way the model actually works. Put another way, a method with high fidelity will assign high relevance to features that, when removed, greatly reduce the DNN's output confidence in the class assignment, while assigning low relevance to features that do not greatly affect the confidence when removed. Fidelity cannot be established axiomatically, and so must be estimated using implementations of the explanation method with the DNN model and dataset under investigation.

In this paper, we investigate the properties of different approaches for measuring explanation method fidelity, focusing on saliency map explanations for image classifiers. We make the distinction between estimating the *fidelity of an explanation method*, and estimating the *fidelity of an individual explanation*. While currently proposed metrics assess the former, it is the latter that we are often interested in — especially if we intend to use individual model outputs for decision making. Inspired by (Adebayo et al. 2018), who introduced a set of "sanity checks" (NB: *not* metrics) for testing saliency methods, we propose a corresponding set of sanity checks for saliency *metrics* based on measures of reli-

---

[*]The first two authors contributed equally.

ability from the psychometric testing literature. We use these checks to show that current metrics used to estimate saliency method fidelity can have high variance and are sensitive to implementation details. We also show that these metrics can be statistically unreliable when considered over individual saliency maps, and that the metrics we examined were not measuring the same underlying notion of fidelity.

Background

Saliency methods

Many methods have been proposed for improving the interpretability of neural networks (NNs). These can broadly be categorized into two types: (1) methods that investigate the NN's internals, and (2) methods that attempt to explain the NN output for individual inputs (Lipton 2016). The former type of approaches are useful to researchers and engineers trying to understand their models, but less useful to a broader audience — those without access to the model structure (e.g. those accessing a model-as-a-service via an API), or users without the relevant knowledge for interpreting information about NN internals. The latter approach of offering easily interpretable, "post-hoc" explanations (Lipton 2016) of outputs for specific inputs has therefore received much attention.

Feature attribution methods — often called saliency maps when the features are pixels in images (Adebayo et al. 2018) — are a popular explanation technique for explaining classification outputs (Montavon, Samek, and Müller 2018). These methods provide a measure of the relevance that each feature/pixel in a given input has to the model's output. Obtaining a correct value for the relevance of each feature is non-trivial: NNs are able to learn non-linear combinations of features that make separating their relative relevance to any particular output difficult. Additionally, many kinds of data have considerable local structure, and NN models are usually developed to take advantage of this. For example, pixels are only meaningful within an image when considered with nearby pixels, a property exploited by convolutional neural networks (CNNs).

Previously proposed methods for creating saliency maps for images have employed a variety of techniques for inferring pixel relevance (Montavon, Samek, and Müller 2018). These have a variety of mathematical formulations that result in different properties for the saliency maps they produce. However, many are based on common theoretical grounds: for example, gradient-based methods rely on back-propagating gradients from output to input units to estimate how changes in the input will affect the output (Selvaraju et al. 2017), while relevance-based methods estimate pixel relevance with reference to a root point in the input space (Montavon et al. 2017). Several authors have explored the theoretical relationships between different methods, finding many commonalities between different approaches (Lundberg and Lee 2017; Ancona et al. 2017; 2018).

Saliency metrics

Explanation evaluation is inherently difficult because many different aspects of an explanation affect its perceived quality. One aspect of an explanation that should be quantifiable is the explanation's *fidelity*: how well it genuinely represents the processing performed by a model on the input to produce the corresponding output. A measure of fidelity for saliency map methods should capture how well the method assigns relevance values to the input pixels.

Despite the large number of saliency methods available, relatively few metrics have been proposed for assessing their fidelity. One of the first approaches, initially outlined by (Bach et al. 2015) for binary images and extended by (Samek et al. 2017) for RGB images, measures the change in classifier output as pixels are sequentially perturbed (flipped in binary images, or set to a different value for RGB images) in order of their relevance as estimated by the saliency method. The classification output should decrease more rapidly for methods that provide more accurate estimates of pixel relevance. Similarly, if the *least* relevant pixels are perturbed first (Arras et al. 2017), then the classification output should change more slowly the more accurate the saliency method. The proposed metric for capturing this is the Area Over the Perturbation Curve (AOPC_M), which can be measured by perturbing pixels based either on the Most Relevant First (MoRF) or the Least Relevant First (LeRF) procedures:

$$\mathrm{AOPC}_M = \frac{1}{L+1} \left\langle \sum_{k=1}^{L} f(\mathbf{x}_M^{(0)}) - f(\mathbf{x}_M^{(k)}) \right\rangle_{p(\mathbf{x})},$$

where M is the pixel deletion procedure (MoRF or LeRF), L is the number of pixel deletion steps, $f(\mathbf{x})$ is the output value of the classifier for input image $\mathbf{x}$ (i.e. the probability assigned to the highest-probability class), $\mathbf{x}_M^{(0)}$ is the input image after 0 perturbation steps (i.e. $\mathbf{x}_M^{(0)} = \mathbf{x}$), $\mathbf{x}_M^{(k)}$ is the input image after k perturbation steps, and $\langle \cdot \rangle_{p(\mathbf{x})}$ denotes the mean over all images in the data set (Samek et al. 2017). Varying L allows the measurement of AOPC_M over different amounts of pixel deletion.

A second metric described by (Alvarez-Melis and Jaakkola 2018) and simply called the "faithfulness", F, similarly relies on removing features (perturbing pixels for images) and measuring the change in classification output. The metric as proposed in (Alvarez-Melis and Jaakkola 2018) differs from AOPC in that it perturbs pixels one-by-one, so that the change in output is only ever caused by a single pixel perturbation. The faithfulness $F_{\mathbf{x}}$ for a single image $\mathbf{x}$ is calculated by taking the Pearson correlation between the relevance R_i assigned to pixel i by the saliency method, and the change in classification output when pixel i is perturbed to create image $\mathbf{x}_i$: $\Delta_i = f(\mathbf{x}) - f(\mathbf{x}_i)$, for all pixels in the image. The saliency method faithfulness, F, is again calculated by taking the average faithfulness over all images in the data set:

$$F = \langle \rho(R, \Delta) \rangle_{p(\mathbf{x})}.$$

A third metric, RemOve And Retrain (ROAR) was proposed by (Hooker et al. 2019). The ROAR procedure calculates saliency maps for each image in the training data,

perturbs the most relevant pixels, then retrains a new model with the same structure and initialization as the original model using the perturbed training data. If the saliency method identified the correct important pixels, the new classifier should exhibit a large reduction in accuracy compared to the original. Unlike the other proposed methods, ROAR does not allow for the measurement of individual saliency map fidelity — only global saliency method fidelity. It thus cannot be used to assess the fidelity of individual saliency maps. Additionally, it requires extensive computational resources to retrain models from scratch after successive perturbation steps. We therefore did not consider ROAR further in our investigations.

Evaluating saliency metrics

A metric should have good *statistical validity*: it should measure the property that it is intended to measure. A true test of saliency metric validity would require knowledge of the ground-truth saliency maps — which we do not have access to, and are what the saliency methods are trying to estimate.[1] However, we can investigate a saliency metric's *reliability* i.e. how well it provides consistent results. While a reliable metric is not necessarily valid, a valid metric *must* be reliable (Davidshofer and Murphy 2005).

To determine reliability, we consider metrics as applied to individual saliency maps, and make use of statistics usually used in the psychometric testing literature. We can think of a set of saliency methods as a battery of psychometric tests administered to an agent (the NN). The tests are scored by a saliency metric. Each input image corresponds to a different rater, who administers the battery of tests. Psychometric test reliability is usually estimated in four separate ways (Peter 1979):

- *Inter-rater reliability*: degree to which different raters agree in their assessments. We use inter-rater reliability to refer to the agreement in the saliency metric scores between input images (raters) across different saliency methods, using the same metric.

- *Inter-method reliability*: degree to which different tests agree in their assessments. We use inter-method reliability to refer to the agreement between different saliency methods (tests) across different input images, using the same metric.

- *Internal consistency reliability*: degree to which different methods that are intended to measure the same concept produce similar scores. We use internal consistency reliability to refer to the agreement between different saliency metrics measuring the same saliency method.

- *Test-retest reliability*: degree to which scores change between test administrations. Test-retest reliability is not relevant for saliency metrics because we are applying deterministic saliency methods to deterministic, fixed models.

Though this analogy is imperfect, it assists in the selection of appropriate statistics to test saliency metric reliability (our sanity checks), as described in the following subsections.

Inter-rater reliability

Using the above analogy, this class of reliability assesses how consistent a saliency metric is in its scores between different images. An initial assessment can be made by measuring the variance of the metric for a single saliency method over all data set images. However, a high variance may be acceptable as the variety of images in the test set may present a broad variety of difficulties for the saliency method. Instead, we can consider how consistent the metric is at ranking the fidelity of different saliency maps, when considered image by image. In other words, does the metric consistently rank some saliency methods higher than others over all the images?

The statistic commonly tested to answer this question is Krippendorf's α, which is defined as

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is the observed disagreement in saliency method ranking between images, and D_e is the disagreement expected by chance, taking into account the number of images and number of methods being ranked (Krippendorff 2004). Full details of how to calculate D_o and D_e are given in (Krippendorff 2004). If $\alpha = 1$, the saliency method ranking between images is totally consistent (i.e. the metric produces the same ranking over methods for every image), while $\alpha = 0$ implies the ranking between images is random (negative values indicate systematic disagreement). If the saliency metrics have a low value of α, it implies that their ranking of saliency map fidelity for future test images is largely unpredictable.

Inter-method reliability

Inter-method reliability assesses whether a saliency metric agrees across different saliency methods. This can be measured by taking the pairwise correlations between the scores of the different saliency methods on the data set images. If the scores of each saliency method fluctuate similarly between images, these correlations will be high, indicating high inter-method reliability. We use Spearman's ρ to measure these pair-wise correlations.

Internal consistency reliability

Internal consistency reliability indicates whether different saliency metrics are capturing the same underlying concept (saliency map fidelity). This can be measured by taking the correlation between the scores produced by different metrics over saliency maps produced by the same saliency method. Again, we use Spearman's ρ to measure this.

---

[1]In early experiments, we attempted to generate ground truth saliency maps by employing rules to generate images usually used in tests of human visual reasoning. However, we were still constrained to using very simple models to obtain these ground truth maps. Ultimately the results of these experiments were of limited use as they were not indicative of the results we obtained using the saliency methods and metrics with more complex models and datasets.

Experiments

Our experiments were designed to measure the reliability of saliency metrics for saliency maps, as outlined in the previous section. We investigated both AOPC_M and faithfulness F.

Methods

Our goal in this paper is to better understand saliency metrics, so instead of testing many saliency methods on many different models and tasks, we focused on a single model and a few saliency methods. This allowed us to explore the properties of the saliency metrics in depth, rather than being distracted by effects of different models and datasets on the metrics.

Model and task We performed all our experiments on a CNN model trained on the CIFAR-10 dataset and classification task (Krizhevsky 2009). We chose CIFAR-10 because it is a well-known image classification set of suitable complexity (10 non-linearly separable classes) whose size (50,000 32×32 RGB training images, 10,000 test images) is not prohibitive to running many experiments in a reasonable amount of time. We trained a standard CNN containing three sequential blocks, with each block consisting of two 2D convolution layers with batch normalization, followed by a max pooling layer. The output of the third block is flattened before connecting to the final classification layer. ReLUs were used for all activations except the classification layer, which used the standard SoftMax activation. The model had all bias terms set to zero, as the inclusion of bias terms can pose difficulties for relevance backpropagation approaches to pixel saliency estimation (Wang, Zhou, and Bilmes 2019; Montavon, Samek, and Müller 2018). The model was trained on 45,000 training samples, with 5,000 held out as a validation set. During training the model was regularized using l_2 weight decay and dropout, and early stopping was used to prevent over-fitting. The resulting model achieved a test set accuracy of 86%.

Saliency methods We chose to compare four different saliency methods and a baseline method (edge detection). Our aim here was not to be exhaustive, but to choose well-known methods from the literature with differing properties and assumptions. The methods chosen were sensitivity analysis (Simonyan, Vedaldi, and Zisserman 2013), gradient⊙input,[2] deep Taylor decomposition (Montavon et al. 2017), and SHAP (Lundberg and Lee 2017) — specifically DeepSHAP, which builds on a connection between SHAP and DeepLIFT (Shrikumar, Greenside, and Kundaje 2017). Each of these produces a saliency map that assigns a relevance value to each pixel in the image, relative to the model's output classification score. Example saliency maps are shown in Figure 1.

Figure 1: Example saliency maps using the four saliency methods plus edge detection, shown for five of the ten CIFAR-10 image classes (top to bottom: dog, frog, horse, ship, truck). Positive relevance shown in red, negative in blue. These images were all correctly classified by our NN with confidence ≥ 0.99. Saliency maps were normalized by their maximum absolute value for visualization.

Some methods only estimate positive relevance i.e. all pixels are considered to have made a positive (or zero) contribution to the output classification, while other methods estimate positive *and* negative relevance. Of the methods we looked at, sensitivity analysis and deep Taylor decomposition estimate positive relevance only, while gradient⊙input and SHAP estimate positive and negative relevance. We use the implementations of gradient (which we convert to sensitivity by taking the channel-wise maximum of the magnitude of the gradient), gradient⊙input, and deep Taylor decomposition provided by the iNNvestigate toolbox (Alber et al. 2019), and the implementation of DeepSHAP available at https://github.com/slundberg/shap/ . Finally, we also compare these methods with saliency maps produced by Sobel edge detection, which acts as a baseline. Edge detection creates maps that are visually similar to those produced several saliency methods, but that do not depend on the NN internals (Adebayo et al. 2018).

Saliency metrics We investigated AOPC with both MoRF and LeRF pixel perturbation order, and faithfulness F. These metrics require pixels to be "perturbed" but for colour images, the correct perturbation function is not obvious. Several different approaches have been proposed in the literature, largely without explicitly stating why that particular function was chosen. We use two different perturbations: replacing the selected pixel with the dataset mean, and replacing the selected pixel with uniformly distributed random RGB values. The former approach effectively sets the pixels to 0 as we standardize inputs to the network, while the latter attempts to destroy the information contained in the pixel, as well as its correlation with surrounding pixels. Some pre-

---

[2]It was noted by (Kindermans et al. 2016; Shrikumar, Greenside, and Kundaje 2017) that gradient⊙input was functionally equivalent to a version of the Layerwise Relevance Propagation method, ϵ-LRP, (Bach et al. 2015; Kindermans et al. 2016) under certain conditions. We use gradient⊙input here as it is computationally simpler, and equivalent to ϵ-LRP for our network.

vious studies have chosen to perturb square pixel-regions rather than individual pixels (Samek et al. 2017); we chose to perturb single pixels as perturbing pixel regions requires assumptions about the spatial scale of the features in the images and the information contained in the saliency maps. We also note that perturbing a single pixel in a CIFAR10 image is similar to perturbing a 9×9 region in an ImageNet (Deng et al. 2009) image as done by e.g. (Samek et al. 2017) (1 pixel is 0.1% of a CIFAR10 image's area; a 9×9 region is 0.16% of an ImageNet image's area assuming standard resizing to 224×224).

For the faithfulness results, we did not perturb every pixel in every image as this would have been computationally prohibitive. Instead we randomly selected 100 pixel IDs and perturbed this same pixel set for every image to obtain an estimate of the faithfulness. For the AOPC results, we also scored a set of random perturbation orderings to obtain a random baseline in addition to the edge detection baseline. We generated 100 random pixel orderings, using this set to obtain a set of 100 AOPC scores for every image. We use the mean of these 100 scores as the random baseline score, with a 95% confidence interval calculated on the empirical distribution of mean scores. Other confidence intervals were estimated using the bootstrap method with 10,000 re-samplings. We performed our experiments on the whole of the CIFAR-10 test set of 10,000 images (1000 per class).

Results

Global saliency metric reliability

Figure 2 shows the two metrics as measured for each saliency method, obtained using both mean perturbation (setting perturbed pixels to the data set mean value) and random RGB perturbation (setting perturbed pixels to uniformly random RGB values). Figure 2a shows AOPC values plotted against number of deletion steps; the top row shows results for MoRF perturbation and the bottom row for LeRF. These plots reveal several details about the reliability of AOPC measures. Considering first MoRF perturbation order (where higher AOPC values indicate better fidelity), gradient⊙input and SHAP are the top-ranked methods for mean and random RGB perturbation, respectively. Deep Taylor decomposition and sensitivity are indistinguishable using mean perturbation, but sensitivity is ranked better than deep Taylor decomposition with random RGB perturbation. This inconsistency indicates that $\mathrm{AOPC_{MoRF}}$ is sensitive to the details of the perturbation function for our CNN. Turning to LeRF perturbation order (where lower AOPC values indicate better fidelity), the saliency methods are ranked consistently between mean and random RGB perturbation. However, the LeRF rankings are opposite to those measured by MoRF: edge detection and deep Taylor decomposition obtain the best scores, with SHAP third and gradient⊙input last. This indicates that $\mathrm{AOPC_{MoRF}}$ and $\mathrm{AOPC_{LeRF}}$ are not measuring the same thing (low internal consistency reliability). A final thing to note regarding the plots in Figure 2a is the hidden variance of the AOPC measures: the 95% confidence intervals for the means are very tight due to the large sample size, but the variance in AOPC scores over in-

put images is large.

Figure 2b shows Faithfulness, F, for each saliency method. In these plots we show the full distribution of F scores measured on each image, as well as the means indicating the global scores for the different methods. The mean values for every method are close to zero, indicating that they all have low faithfulness, with edge detection consistently the lowest. Again, the perturbation method affects the metric, changing the distribution shapes, mean values, and method rankings.

Local saliency metric reliability

Table 1 lists the Krippendorf α statistics for the saliency metrics (with varying numbers of deleted pixels L for AOPC). These are calculated on the image-wise rankings of saliency maps, and test "inter-rater reliability" as described above. Low α values indicate that the saliency method rankings on different images are inconsistent. The left two columns of table 1 show α values when the baseline edge detection method is included in the ranking. This baseline is ranked more consistently (low rankings for $\mathrm{AOPC_{MoRF}}$ and F, high rankings for $\mathrm{AOPC_{LeRF}}$) than the true saliency methods over all data set images, producing higher α values than when it is excluded from the rankings (the right-hand two columns).

Perturbing with random RGB values reduces α for Fidelity and $\mathrm{AOPC_{MoRF}}$ compared with mean perturbation, which may be due to the increased stochasticity in the perturbations from the random colour choices. However, it is difficult to understand why $\mathrm{AOPC_{LeRF}}$ at smaller numbers of perturbation steps ($L = 20$ and $L = 40$) produce greater α values with random RGB perturbation that with mean perturbation. A "low" α value is not strictly defined, but $\alpha < 0.65$ are often considered to indicate unreliability inter-rater reliability. The largest value in the table, 0.48 for $\mathrm{AOPC_{MoRF}}$ at $L = 100$ using mean perturbation and including edge detection in the ranking, is substantially less than 0.65, indicating low inter-rater reliability whatever the specifics of the metric. We can conclude from this that the metrics will not produce consistent rankings of saliency maps when applied to new test images.

We made a similar set of measurements looking at the inter-method correlation of metric scores over data set images, taking the mean of the pair-wise Spearman's ρ correlation between saliency methods. As with the previous results, the presence of the baseline edge detector — which should not correlate well with other methods given that it does not assess the model — reduces the mean pairwise correlation. The lowest mean pairwise correlation was 0.13 (for Faithfulness, random RGB perturbation, edge detection excluded) and the highest 0.67 (for $\mathrm{AOPC_{MoRF}}$, $L = 100$, mean perturbation, edge detection excluded). Correlations for $\mathrm{AOPC_{MoRF}}$ were consistently highest, ranging between 0.47 and 0.67, while Faithfulness showed the lowest correlation values of 0.13 to 0.18 (both excluding edge detection). Changes in perturbation method appear to affect intermethod reliability much less than inter-rater reliability.

Finally, we assess internal consistency reliability by measuring the correlation between different metrics over the data
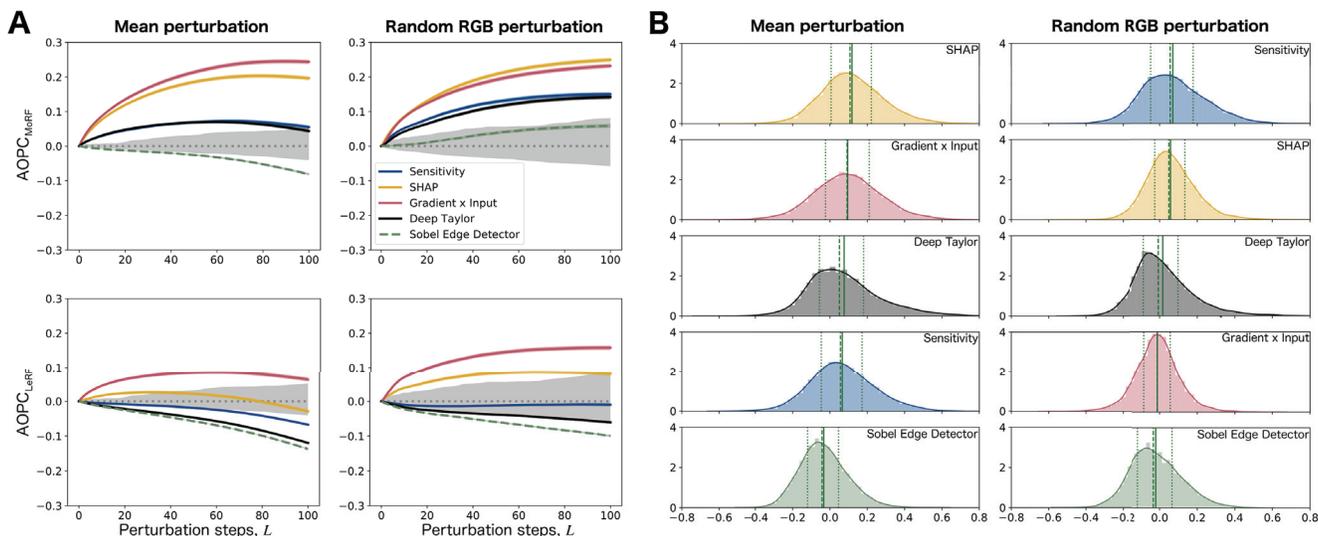
Figure 2: Metrics for saliency methods. **(a)** AOPC scores for each saliency method, measured from $L = 0$ to $L = 100$ perturbation steps. Scores adjusted by subtracting the mean AOPC for random deletion (dotted line with grey shading). *Left column*: results using mean perturbation. *Right colum*: results using random RGB perturbation. *Top row:* Results for MoRF pixel order. Higher $\text{AOPC}_{\text{MoRF}}$ is better. *Bottom row*: results for LeRF pixel order. Lower $\text{AOPC}_{\text{LeRF}}$ is better. 95% confidence intervals shaded (very tight, so barely visible except for the random baseline). **(b)** Faithfulness (F) scores for each saliency method. Distribution over all images is show as a histogram and density plot. Mean shown as vertical solid line; median is vertical dashed line; upper and lower quartiles are vertical dotted lines. Rows are ordered by metric rank (highest mean F top row, to lowest mean F bottom row. *Left column*: results using mean perturbation. *Right column*: results using random RGB perturbation.

	Mean perturb, with edge detection	Random perturb, with edge detection	Mean perturb, excluding edge detection	Random perturb, excluding edge detection
Faithfulness, F	0.18 (0.17—0.20)	0.10 (0.09—0.12)	0.04 (0.03—0.05)	0.09 (0.08—0.10)
$\text{AOPC}_{\text{MoRF}}$, L=20	0.29 (0.27—0.31)	0.15 (0.14—0.17)	0.19 (0.17—0.20)	0.07 (0.06—0.08)
$\text{AOPC}_{\text{MoRF}}$, L=40	0.38 (0.36—0.39)	0.18 (0.17—0.20)	0.27 (0.25—0.28)	0.09 (0.08—0.11)
$\text{AOPC}_{\text{MoRF}}$, L=60	0.42 (0.40—0.44)	0.20 (0.18—0.21)	0.31 (0.29—0.33)	0.11 (0.10—0.12)
$\text{AOPC}_{\text{MoRF}}$, L=80	0.46 (0.44—0.48)	0.21 (0.20—0.23)	0.34 (0.32—0.36)	0.12 (0.11—0.14)
$\text{AOPC}_{\text{MoRF}}$, L=100	0.48 (0.46—0.50)	0.22 (0.21—0.24)	0.37 (0.35—0.39)	0.13 (0.12—0.15)
$\text{AOPC}_{\text{LeRF}}$, L=20	0.05 (0.04—0.06)	0.13 (0.11—0.15)	0.06 (0.05—0.07)	0.15 (0.13—0.17)
$\text{AOPC}_{\text{LeRF}}$, L=40	0.11 (0.09—0.12)	0.18 (0.16—0.20)	0.11 (0.09—0.12)	0.20 (0.18—0.22)
$\text{AOPC}_{\text{LeRF}}$, L=60	0.16 (0.14—0.17)	0.22 (0.20—0.24)	0.15 (0.14—0.17)	0.23 (0.21—0.25)
$\text{AOPC}_{\text{LeRF}}$, L=80	0.20 (0.18—0.22)	0.25 (0.23—0.27)	0.19 (0.17—0.21)	0.26 (0.23—0.28)
$\text{AOPC}_{\text{LeRF}}$, L=100	0.24 (0.22—0.25)	0.28 (0.26—0.30)	0.23 (0.21—0.25)	0.28 (0.26—0.30)

Table 1: Krippendorf's α under different perturbation functions for the saliency metrics, measured using saliency method ranking across images. Numbers in brackets are 99.9% confidence intervals estimated with 10,000 bootstrap samples.

	Sensitivity	gradient⊙input	SHAP	Deep Taylor	Edge detection
Faithfulness, F vs $\text{AOPC}_{\text{MoRF}}$,	0.34	0.11	0.00	0.24	0.17
Faithfulness vs $\text{AOPC}_{\text{LeRF}}$,	-0.11	-0.14	-0.09	-0.22	-0.09
$\text{AOPC}_{\text{MoRF}}$ vs $\text{AOPC}_{\text{LeRF}}$	0.28	0.53	0.77	0.16	0.11
$\text{AOPC}_{\text{MoRF}}$, random RGB perturb vs $\text{AOPC}_{\text{MoRF}}$, mean perturb	0.62	0.71	0.77	0.58	0.57

Table 2: Spearman correlations between pairs of metrics, measured over all images, for each saliency method. AOPC was taken at $L = 100$ perturbation steps. Unless otherwise stated, mean perturbation was used.

set images, for each saliency method. A selection of pairs of correlations for different metrics is shown in table 2. The correlation values indicate that the metrics are, in general, not measuring the same underlying quantity across the different saliency methods. Faithfulness is only weakly correlated with $\text{AOPC}_{\text{MoRF}}$, and in fact slightly anti-correlated with $\text{AOPC}_{\text{LeRF}}$ across all saliency methods. The correlation between $\text{AOPC}_{\text{MoRF}}$ and $\text{AOPC}_{\text{LeRF}}$ is highly variable across saliency methods, while the most consistent correlations across methods are between $\text{AOPC}_{\text{MoRF}}$ applied with random RGB versus mean perturbation. This is to be expected as both methods perturb the same pixels — though there is still reasonably high variability even in this case (from $\rho = 0.58$ for deep Taylor decomposition to $\rho = 0.77$ for SHAP).

Discussion

We investigated the reliability of several saliency metrics that have been proposed to measure saliency method fidelity. To do this, we applied the metrics to a single model and task, assessing the reliability of the metrics in relation to this model. It is important to note that our results and conclusions are not generally applicable statements about the consistency of saliency metrics — they are an illustration that such metrics *can* produce unreliable and inconsistent results. Users should be wary of these properties when applying saliency metrics, and we suggest they apply the above sanity checks to any metrics in the context of their particular data and model.

Our main findings can be summarized thus:

- *Global saliency metrics had high variance*: The distribution of scores over all test images for each saliency metric was broad. Past work has only reported mean metric values (Arras et al. 2017; Samek et al. 2017; Bach et al. 2015; Montavon, Samek, and Müller 2018), ignoring this variance.

- *Saliency metrics were sensitive to the specifics of their implementation*: we observed that both the faithfulness and AOPC scores, and the ranking of the saliency methods according to these scores, were affected by the specifics

of the perturbation method (in our case, mean vs random RGB perturbation).

- *Saliency maps from different saliency methods were ranked inconsistently image-by-image*: each metric exhibited low inter-rater reliability, meaning that rankings of saliency method fidelity considered image-by-image were highly inconsistent. This suggests that global saliency metric scores provide an unreliable indication of the fidelity of different saliency methods on future test images.

- *The internal consistency of different metrics that all attempt to measure fidelity was low*: the correlation between different metric scores over saliency maps is generally low, and highly variable. This means that the different metrics we tested were not necessarily measuring the same underlying concept.

The above findings point to the great difficulty in measuring saliency map fidelity when we do not have access to ground truth saliency maps. In particular, it is very hard to disentangle the sources of variance and inconsistency in the metric scores, as model, saliency method and metric are tightly coupled. The specifics of the pixel perturbation method are clearly important as demonstrated above: however carefully implemented, these have the potential to move the image off the data manifold learned by the NN, thus reducing the validity of any measurements. (Montavon, Samek, and Müller 2018) recommend choosing a perturbation method that is less likely to move the image off-manifold, but it is not currently possible to know what method will adhere to this requirement a priori.

Another notable observation was the reversal in global saliency method fidelity rankings between $\text{AOPC}_{\text{MoRF}}$ and $\text{AOPC}_{\text{LeRF}}$. Two effects appeared to contribute to this. First, LeRF perturbation order apparently favors methods that are more edge-detection like: edge detection and deep Taylor decomposition perform best with LeRF. These methods assign low relevance to large, contiguous blocks of color, and setting these regions to the dataset mean removes little discriminative information from the image. Equally, mean-perturbing all pixels along an edge does not remove the edge, it merely changes its contrast. The information

in high-relevance pixels is therefore not fully removed by mean perturbation, as illustrated by deep Taylor decomposition's poor performance under MoRF. Also note that edge detection and deep Taylor decomposition's $\text{AOPC}_{\text{MoRF}}$ scores with random RGB perturbation are better than with mean perturbation, as random perturbation removes edge information more successfully. Second, gradient⊙input and SHAP — which produce both positive and negative relevance values — perform worse than random under LeRF (when the most negative pixels are perturbed first, rather than the pixels with the lowest absolute values) but do well under MoRF. We noted that the model was very confident on most inputs (59% of test images classified with confidence ≥ 0.99) despite the presence of negative relevance pixels in gradient⊙input and SHAP saliency maps. When we considered only the 1600 images classified with confidence ≤ 0.8, LeRF scores were better than random for the methods producing negative relevance. This suggests negative relevance in saliency maps for high-confidence classifications contribute to poor LeRF scores.

Further insights into the underlying causes of these results could be gleaned by considering the metric results on each image class individually. We found that all metric scores were dependent on image class; e.g. saliency maps of frog images produced generally better $\text{AOPC}_{\text{MoRF}}$ scores, worse $\text{AOPC}_{\text{LeRF}}$ scores, and better F scores than other classes. The reasons for this are yet to be fully examined, but will depend on the systematic differences in the structure of the higher-level features in different classes, how the classifier has learned to represent these features, and how well the saliency maps capture these representations. We leave a fuller analysis along these lines as future work.

Conclusion

Inspired by (Adebayo et al. 2018)'s sanity checks for saliency maps, we investigated a set of sanity checks for saliency *metrics*. To be useful, metrics should have high statistical validity. As this is not possible to establish without (unavailable) ground truth references, we can at least investigate metric *reliability*. We illustrated the application of different measures of reliability to saliency metrics, specifically inter-rater reliability, inter-method reliability, and internal consistency. These sanity checks showed that current metrics were unreliable at measuring saliency map fidelity for our model. Due to the tight coupling of model, saliency method and saliency metric, it will be very challenging to disentangle the causes of this unreliability.

Based on our results, we propose the following recommendations for future metric development:

- New metrics should be compared directly with previous metrics purporting to measure the same or related quantities, and their differences made explicit (preferably quantitatively).

- If the metric could be implemented in different ways (e.g. mean versus random RGB pixel perturbation), the properties of these metric variants should be analyzed under different conditions to establish the effects of these different implementations.

- Metrics should be analyzed for how they might be "tricked" in different contexts (e.g. edge detection giving the best $\text{AOPC}_{\text{LeRF}}$ scores)

- Metric developers should encourage users of their metric to investigate and understand the sources of variance in the metric scores, and how this affects their decisions about what saliency methods to choose for their particular model.

We also suggest that developers of saliency methods should not rely on a single metric to test their method's fidelity, as the results could be misleading. Instead, they should employ several metrics and attempt to understand sources of unreliability for their method on a variety of datasets and models.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *NeurIPS*, 9525–9536.

Alber, M.; Lapuschkin, S.; Seegerer, P.; Hägele, M.; Schütt, K. T.; Montavon, G.; Samek, W.; Müller, K.-R.; Dähne, S.; and Kindermans, P.-J. 2019. iNNvestigate Neural Networks! *Journal of Machine Learning Research* 20(93):1–8.

Alvarez-Melis, D., and Jaakkola, T. S. 2018. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, 7786–7795.

Ancona, M.; Ceolini, E.; Öztireli, A. C.; and Gross, M. H. 2017. A unified view of gradient-based attribution methods for deep neural networks. *arXiv:1711.06104*.

Ancona, M.; Ceolini, E.; Öztireli, A. C.; and Gross, M. H. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*.

Arras, L.; Montavon, G.; Müller, K.-R.; and Samek, W. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 159–168.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Muller, K.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10(7).

Davidshofer, K., and Murphy, C. O. 2005. Psychological testing: principles and applications.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Hooker, S.; Erhan, D.; Kindermans, P.; and Kim, B. 2019. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, 9734–9745.

Kindermans, P.-J.; Schütt, K.; Müller, K.-R.; and Dähne, S. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. In *NeurIPS Workshop on Interpretable Machine Learning in Complex Systems*.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2017. The (Un)reliability of saliency methods. In *NeurIPS workshop on Explaining and Visualizing Deep Learning*.

Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. *MSc Thesis, University of Toronto*.

Lipton, Z. C. 2016. The mythos of model interpretability. *arXiv:1606.03490*.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *NeurIPS*, 4765–4774.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65:211–222.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73:1–15.

Montavon, G. 2019. *Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison*. Cham: Springer International Publishing. 253–265.

Peter, J. P. 1979. Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research* 16(1):6–17.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *SIGKDD*, 1135–1144. ACM.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K.-R. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. *arXiv:1704.02685*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *ICML*.

Wang, S.; Zhou, T.; and Bilmes, J. 2019. Bias also matters: Bias attribution for deep neural network explanation. In *ICML*, 6659–6667.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, 818–833. Springer, Cham.